Task 1. Data explorations (30%)

Dog breeds

studentNumber

The dataset for this task contains information on different dog breeds (`dog_breed`). The variables that you require for this assignment are:

- `category`: breed category
- `lifetime_costs`: expected lifetime costs in dollars
- `suitability_for_children`: suitability for children (1 = high, 2 = medium, 3 = low)
- `longevity`: mean life expectancy in years
- `popularity_inUS_ranking`: popularity in the US, 1 = most popular, higher values indicates lower popularity

```r
# Load packages

library(tidyverse)
library(foreign)

# Import data:
dogbreeds <- read_csv2("DogBreeds_selected.csv")
```

Template file and submission:

Add your code to the provided template file. Write reproducible and readable code to make the plots for both exercise a and b and make sure that the plots are visible in your output file (.html/.pdf). Keep the data stored in the subfolder so your .Rmd file can reach it.

For more submission instructions, please see the general instructions of this Graded Assignment.

a) Data descriptives - distributions per variable

Create suitable data visualizations to show the individual distributions of the breeds' mean life expectancy (`longevity`), suitability for children (`suitability_for_children`), categories (`category`) and expected lifetime costs (`lifetime_costs`). For each plot, only use the data that is available, i.e. exclude missing observations.

```r
library(RColorBrewer)
library(ggmosaic)
library(see)



summary(dogbreeds)
```

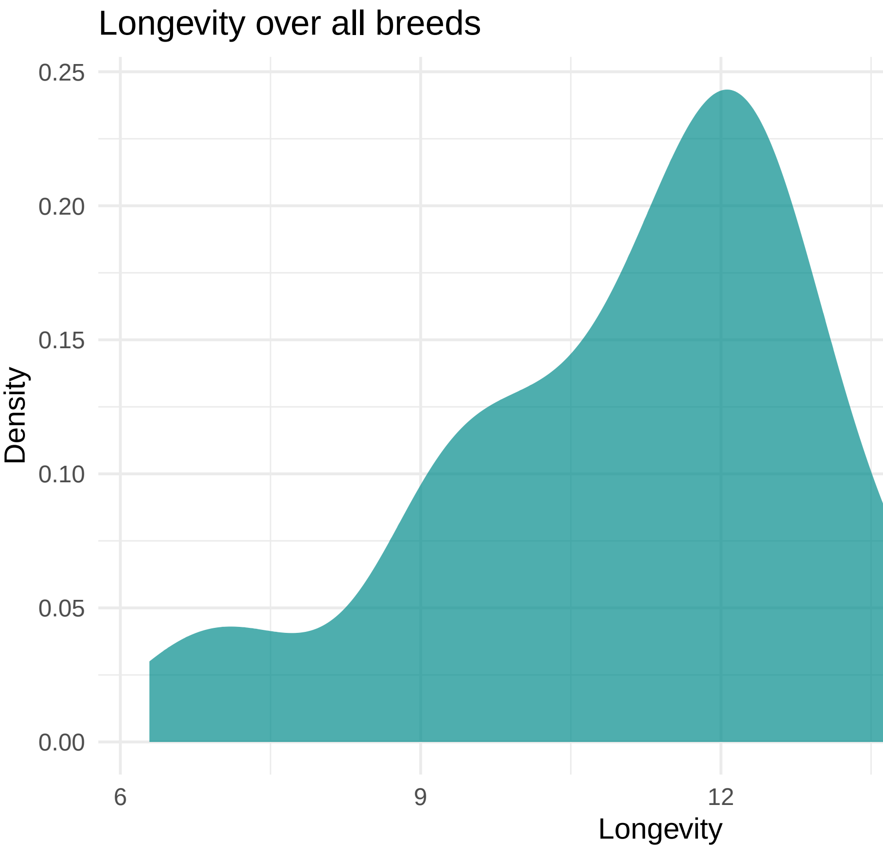
```
##   dog_breed           category          datadog_score   popularity_inUS_ranking
##  Length:174         Length:174         Min.   :0.990   Min.   : 1.0
##  Class :character   Class :character   1st Qu.:2.185   1st Qu.:22.5
##  Mode  :character   Mode  :character   Median :2.710   Median :44.0
##                                        Mean   :2.604   Mean   :44.0
##                                         3rd Qu.:3.035   3rd Qu.:65.5  
##                                         Max.   :3.640   Max.   :87.0  
##                                         NA's   :87      NA's   :87    
##    longevity     grooming_required suitability_for_children size_category
##  Min.   : 6.29   Min.   :1.000     Min.   :1.000            Length:174   
##  1st Qu.: 9.70   1st Qu.:2.000     1st Qu.:1.000            Class :character
##  Median :11.29   Median :2.000     Median :1.000            Mode  :character
##  Mean   :10.96   Mean   :1.804     Mean   :1.491                         
##  3rd Qu.:12.37   3rd Qu.:2.000     3rd Qu.:2.000                         
##  Max.   :16.50   Max.   :3.000     Max.   :3.000                         
##  NA's   :39      NA's   :62        NA's   :62                            
##  intelligence_category lifetime_costs  average_purchase_price
##  Length:174            Min.   :11100   Min.   : 283.0        
##  Class :character      1st Qu.:15386   1st Qu.: 587.2        
##  Mode  :character      Median :17336   Median : 795.0        
##                        Mean   :17069   Mean   : 876.8        
##                        3rd Qu.:18861   3rd Qu.:1042.2        
##                        Max.   :22640   Max.   :3460.0        
##                        NA's   :83      NA's   :28            
##  price_category    
##  Length:174        
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
##                    
```

```r
cleaned <- na.omit(dogbreeds)
```

Note: These plots are expected to be self-contained (i.e. readers should be able to understand them without extra explanation) and to obey the principles of good graphics, but they are not meant to be formal presentation graphics. For example, you are not expected to use additional information to make the plot information rich. The focus is on uncovering the distributions of the variables.

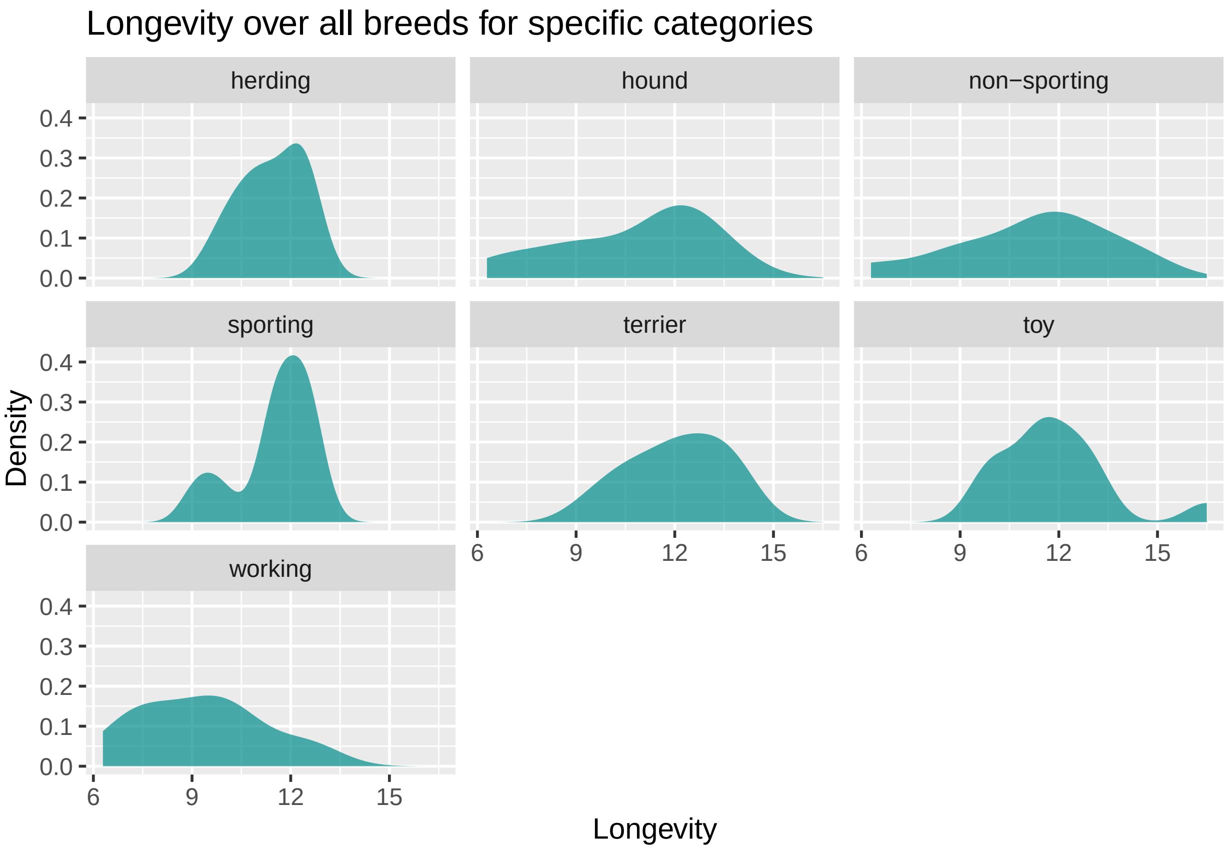
```r
# your code here

ggplot(data = cleaned, aes(x = longevity)) + geom_density(fill="darkcyan",
                                                                         alpha = 0.7,
                                                                         color=NA) +
  labs(x= "Longevity", y ="Density") +ggtitle("Longevity over all breeds")+ theme_minimal()
```

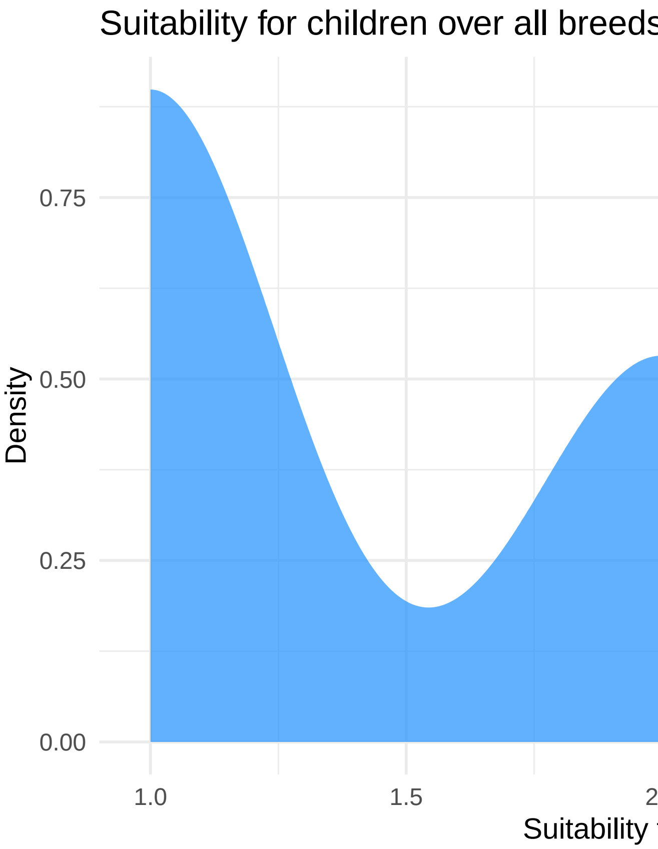


Explore mean life expectancy (`longevity`)

```
ggplot(data = cleaned, aes(x = longevity)) + geom_density(fill="darkcyan",
                                                                         alpha = 0.7,
                                                                         color=NA) +
  labs(x= "Longevity", y ="Density") +ggtitle("Longevity over all breeds for specific categories") + fa
```

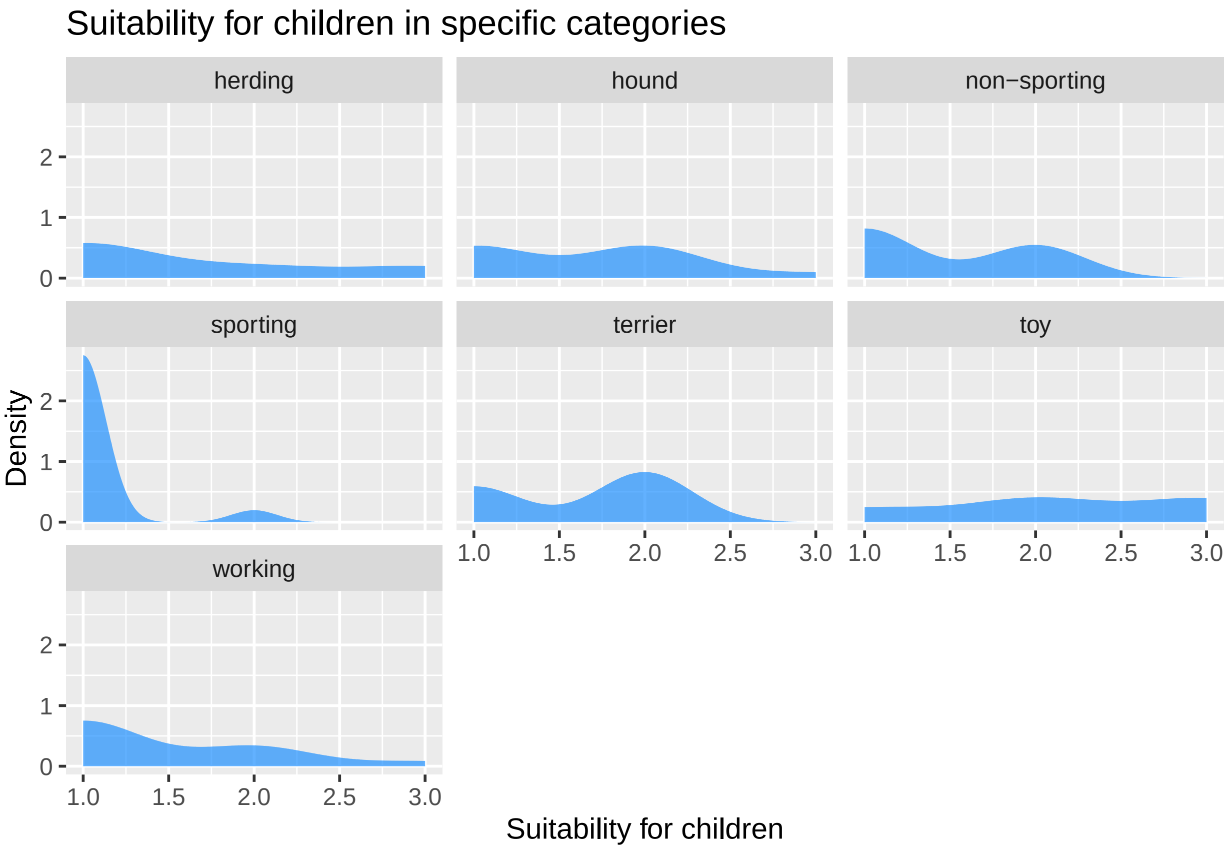


```
# your code here
ggplot(data = cleaned, aes(x = suitability_for_children)) + geom_density(fill="dodgerblue",
                                                                         alpha = 0.7,
                                                                         color=NA) +
  labs(x= "Suitability for children", y ="Density") +ggtitle("Suitability for children over all breeds")
```



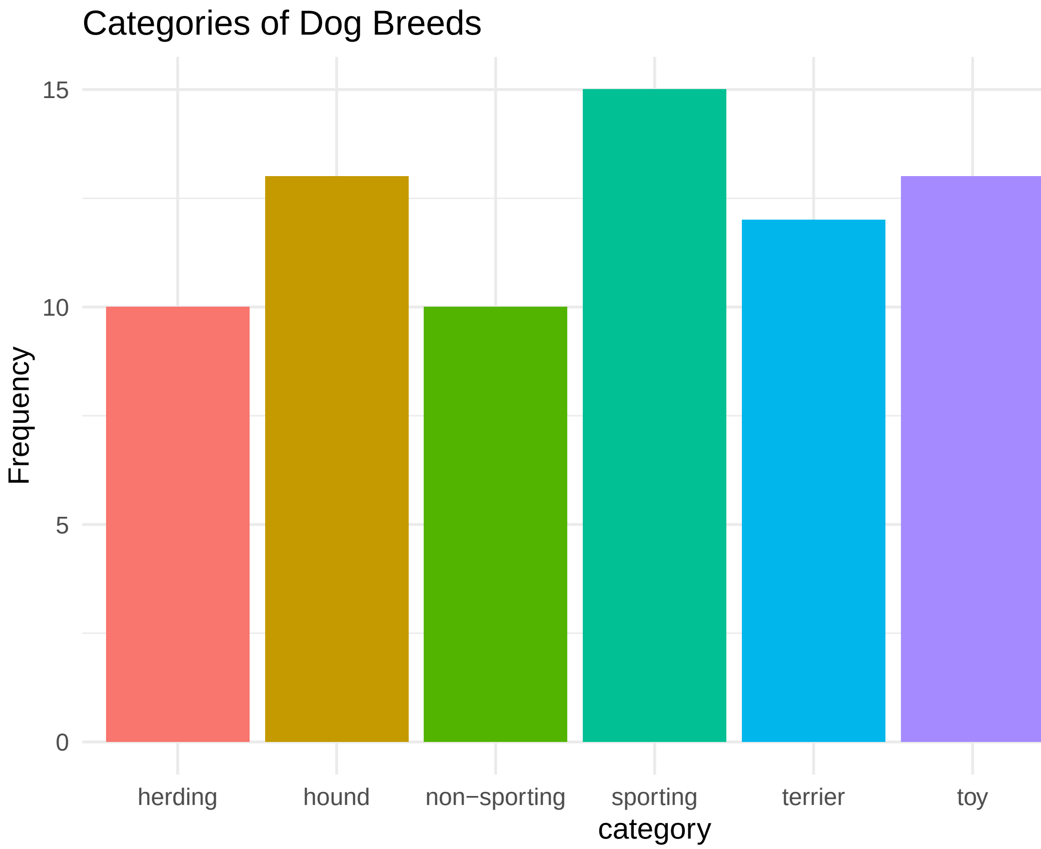
Explore suitability for children (`suitability_for_children`)

```
ggplot(data = cleaned, aes(x = suitability_for_children)) + geom_density(fill="dodgerblue",
                                                                         alpha = 0.7,
                                                                         color=NA) +
  labs(x= "Suitability for children", y ="Density") +ggtitle("Suitability for children in specific cate
```



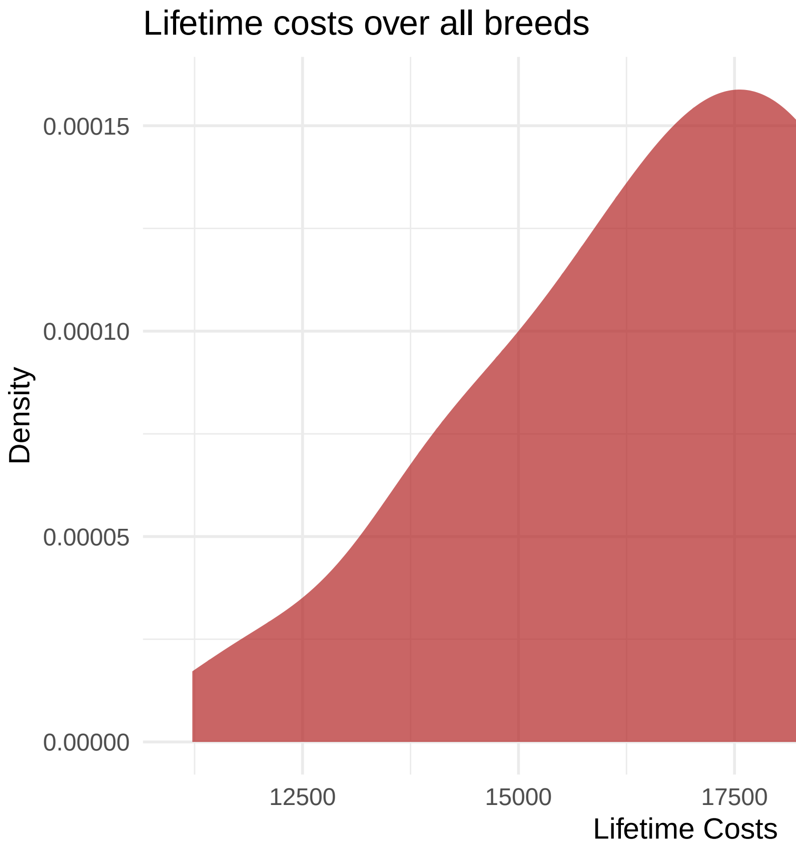
```
# your code here

ggplot(data = cleaned,
       mapping = aes(
         x = category,
         fill = category
       )) + geom_bar(stat="count") +theme_minimal() + theme(legend.position="None") + ggtitle("Categori
```



Explore categories (`category`)

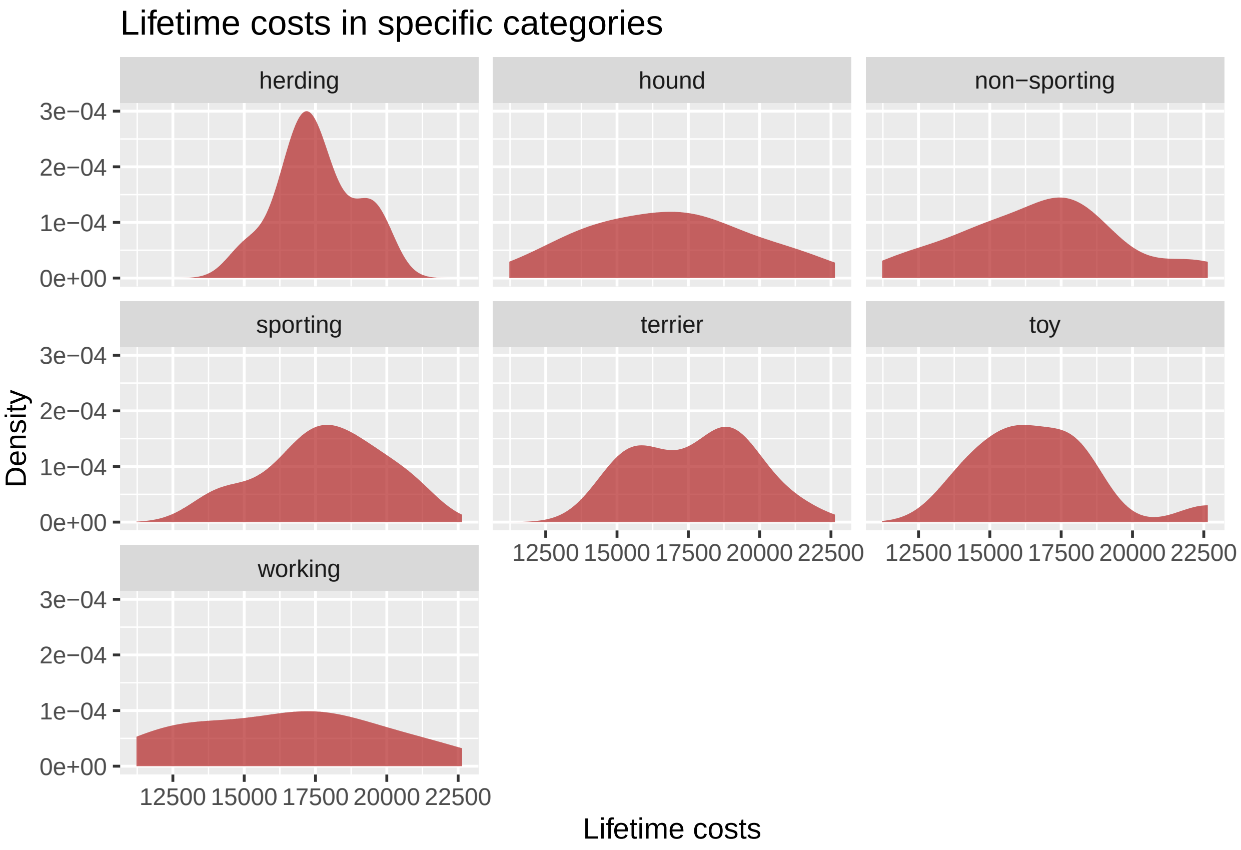
```
ggplot(data = cleaned, aes(x = lifetime_costs)) + geom_density(fill="firebrick",alpha = 0.7,
                                                                         color=NA) +
  labs(x= "Lifetime Costs", y ="Density") +ggtitle("Lifetime costs over all breeds") + theme_minimal()
```



Explore expected lifetime costs (`lifetime_costs`)

```
# your code here

ggplot(data = cleaned, aes(x = lifetime_costs)) + geom_density(fill="firebrick",
                                                                         alpha = 0.7,
                                                                         color=NA) +
  labs(x= "Lifetime costs", y ="Density") +ggtitle("Lifetime costs in specific categories") + facet_wra
```



b) Data descriptives - relationships among 3 variables

Assume that you are studying reasons why some dog breeds are more popular than others. Your main hypothesis is that a dog breed's popularity in the US (`popularity_inUS_ranking`) depends on the expected lifetime costs of the breed (`lifetime_costs`). Additionally, you think that a breed's suitability for children (`suitability_for_children`) might be relevant.

Make a visualisation that focuses mostly on exploring your main hypothesis that popularity depends on lifetime costs, but also gives an indication of whether the breed's suitability for children has an influence.

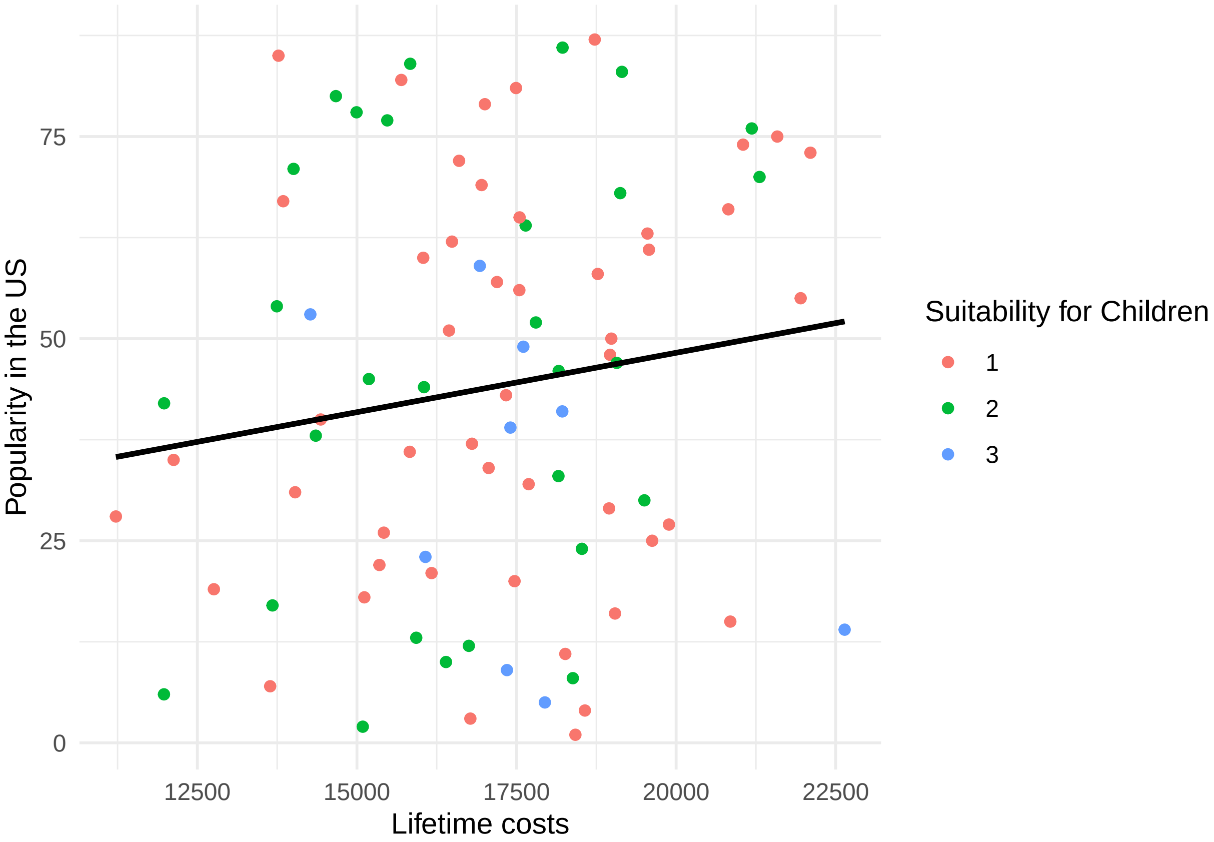
Note: This plot is expected to be self-contained (i.e. readers should be able to understand them without extra explanation) and to obey the principles of good graphics, but it is not meant to be formal presentation graphics. For example, you are not expected to use additional information to make the plot information rich. The focus is on uncovering the relationship among the variables.

```r
# your code here

ggplot(data = cleaned,
       mapping = aes(
         x = lifetime_costs,
         y = popularity_inUS_ranking

       )) + geom_point( aes(color = as.factor(suitability_for_children)))  +
  geom_smooth(method="lm", se=FALSE, color="black") + theme_minimal() +
  labs(x="Lifetime costs", y="Popularity in the US", color="Suitability for Children")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



From the graph above it is safe to say that while there is a positive relationship between lifetime costs and popularity in the US, there is very little effect of suitability for children on popularity (seen by the points both above and below).